Chapter 3
Robustness for Adversarial Risk Analysis

David Ríos Insua, Fabrizio Ruggeri, Cesar Alfaro, and Javier Gomez

Abstract Adversarial Risk Analysis is an emergent paradigm for supporting a decision maker who faces adversaries in problems in which the consequences are random and depend on the actions of all participating agents. In this chapter, we outline a framework for robust analysis methods in Adversarial Risk Analysis. Our discussion focuses on security applications.

3.1 Introduction

Large scale terrorist events like S-11 led to huge security investments. In turn, this has promoted many modeling efforts to support how to efficiently allocate such resources. Parnell et al. [15] provided an in-depth review for the US National Academy of Sciences on bio-terrorism assessment, concluding, among other things, that traditional risk analysis tools, like event trees, are not adequate in this application area for not accounting for adversarial intentionality; the critical and, in many contexts, doubtful common knowledge assumptions of game theoretic approaches;

D. Ríos Insua (✉)
Instituto de Ciencias Matemáticas, Consejo Superior de Investigaciones Científicas, 28049 Madrid, Spain
e-mail: david.rios@icmat.es

F. Ruggeri
Consiglio Nazionale delle Ricerche, Istituto di Matematica Applicata e Tecnologie Informatiche, Via Bassini 15, 20133 Milano, Italy
e-mail: fabrizio@mi.imati.cnr.it

C. Alfaro • J. Gomez
Department of Statistics and Operations Research, Rey Juan Carlos University, Fuenlabrada, 28943 Madrid, Spain
e-mail: cesar.alfaro@urjc.es; javier.gomez@urjc.es

M. Doumpos et al. (eds.), *Robustness Analysis in Decision Aiding, Optimization, and Analytics*, International Series in Operations Research & Management Science 241, DOI 10.1007/978-3-319-33121-8_3

and, finally, the problems of decision analytic based approaches in forecasting adversarial actions. Merrick and Parnell [12] reviewed numerous approaches in this research area, commenting favorably on Adversarial Risk Analysis (ARA), which is a framework to manage risks derived from actions of intelligent adversaries, see [20] or [1].

ARA aims at providing one-sided prescriptive support to one of the intervening agents, the Defender (D, she), based on a subjective expected utility model treating the adversary's decisions as uncertainties. To do so, we model the adversary's (A, Attacker, he) decision making problem and, assuming that he is an expected utility maximizer, try to assess his probabilities and utilities. We can consequently forecast his optimal action. However, our uncertainty about the adversary's probabilities and utilities is propagated to his decision, leading to a random optimal adversary decision which provides us with the required distribution over the Attacker's decision. Sometimes such assessments may lead to a hierarchy of nested decision problems, as described in [17], similar to the concept of level-k thinking, see [24]. In contrast with game theoretic approaches, we do not assume the standard, but unrealistic, common knowledge hypothesis, see [5], according to which the agents share information about their utilities and probabilities.

A critical issue in ARA is elicitation. As in any subjective Bayesian analysis, one needs personal probabilities over the parameters in the problem. Obtaining them is not easy and we need to cope with many biases, see e.g., [14]. This is aggravated in our context because of the involved strategic considerations. Nau [13] as well as Wang and Bier [26] provide discussions of elicitation in the context of adversarial situations.

The practical difficulty of elicitation raises the question of robustness. One wants an ARA to be robust to the elicited probabilities and utilities, the model entertained and, when available, the data. A good way forward is sensitivity analysis. The above mentioned review by Parnell [15] recommends it, and Von Winterfeldt and O'Sullivan [25] perform a systematic sensitivity analysis with respect to elicited probabilities in an event tree concerning MANPADS. A different approach is taken by Kardes [9], who considers robust stochastic games.

Robust Bayesian analysis facilitates finding the entire set of posterior distributions for a parameter when the prior lies within a class of distributions. The results are typically expressed in terms of upper and lower bounds on probabilities and expected utilities. Berger et al. [2] review this methodology which has yet to be used in ARA. The only direct application is given by McLay et al. [11], who point the way towards a principled means to incorporate robustness into ARA. They consider a level-k thinking analysis of the sequential Defend-Attack game in which the Attacker imperfectly observes the decision made by the Defender. The game is modeled through an information structure comprising several signals and, conditional on the defense choice, there is a specified distribution over the signals, a model initially proposed by Rothschild et al. [22]. Robustification occurs by setting upper and lower bounds over parameters for which distributions must be elicited, and then calculating the outcome under the worst case combination of upper and lower values.

This chapter provides a complete outline of the role of robust methods in ARA. After introducing basic notions in Bayesian robustness, we first describe the robust ARA approach for sequential games and, then, for simultaneous games. In both cases, we start by computing the game theoretic solution. We apply robust concepts to assess such solution. If it is not robust, we use the ARA approach to find an alternative solution. Again, we criticize it through robust ideas. If the solution is still unstable, we may appeal to conventional robust concepts, such as the γ maximin. We illustrate the ideas with a simple numerical example concerning routing security.

3.2 Bayesian Robustness

We present here the basic ideas on Bayesian robustness. We refer to Ríos Insua and Ruggeri [18] for an in-depth overview. In the Bayesian approach to inference, prediction and decision making, the interest frequently lies on the behavior of the posterior distribution on a parameter θ obtained by combining experimental evidence provided by the likelihood and expert knowledge expressed through the prior distribution, via Bayes theorem. This is used to compute posterior (and predictive) expectations of functions $g(\theta)$ which typically will be set indicators, powers or utility functions, providing, respectively, set probabilities, moments and expected utilities. The robust Bayesian approach stems from the practical difficulty of specifying a unique prior distribution and/or a unique utility function, corresponding, respectively, to the expert's beliefs and the decision maker's preferences. Therefore, classes of priors and/or utilities are entertained and the consequences of different possible choices of such pairs are evaluated through synthetic indices which determine whether the quantity of interest is subject to small or large variations when changing the prior/utility, i.e. whether there is robustness or not.

In accordance with the content of this chapter, we shall consider utilities u in a class $\mathscr{U}$ and probability measures p in a class $\mathscr{P}$ (without distinguishing whether they are priors or posteriors). We suppose that the probability measure p has a density $p(s)$ over the states s, and the utility function has the form $u(d,s)$, where d is an action (decision) in the feasible set $\mathscr{D}$. We are interested in computing the expected utilities $\psi_{up}(d) = \int u(d,s)p(s)ds$ for various alternatives d and the feasible alternative $d^*_{up} \in \mathscr{D}$ maximizing expected utility, given such choice u and p.

In a robust context, the interest would typically be in the ranges that relevant quantities span when p and u vary in the class, e.g. the range of the expected utility for a certain alternative d

$$\rho_\psi(d) = \sup_{p\in\mathscr{P},u\in\mathscr{U}} \psi_{up}(d) - \inf_{p\in\mathscr{P},u\in\mathscr{U}} \psi_{up}(d),$$

or the distance between the optimal alternative and a reference alternative d^*

$$\rho_d = \sup_{p\in\mathscr{P},u\in\mathscr{U}} e(d^*_{up}, d^*),$$

for some distance e. Looking at ρ_d, we claim that there is robustness if its value is *small* with respect to the entertained problem and the decision maker's perception. In this case, essentially any p and u, and the corresponding d^*_{up}, may be used for decision making purposes. Otherwise, efforts are required to get smaller classes until either robustness can be achieved or no further refinement is possible.

In the latter case, some criterion could be introduced to choose a pair (p,u) and the corresponding d^*_{up}. A possible choice for a decision could be the minimum regret decision,

$$\hat{d} = \text{argmin}_{d\in\mathscr{D}} \max_{p\in\mathscr{P},u\in\mathscr{U}} \left[\psi_{up}(d^*_{up}) - \psi_{up}(d)\right].$$

For a related discussion see [19]. In particular, the decision $\hat{d}$ is conservative in the sense that it protects against the worst loss in expected utility when replacing an optimal decision d^*_{up} by another one.

3.3 Sequential Games

We start by considering sequential games: one agent first makes her decision and, then, the other agent implements his alternative. As an example, imagine a case in which a company deploys their cybersecurity countermeasures and then, observing them, a hacker decides whether he launches an attack or not towards such company.

Specifically, we consider a Defend-Attack situation in which a Defender chooses a defense $d \in \mathscr{D}$ and, then, the Attacker, having observed the defense, chooses his attack $a \in \mathscr{A}$. The corresponding bi-agent influence diagram is shown in Fig. 3.1. An arc reflects that the Defender's choice is observed by the Attacker. The consequences for both players depend on the success s of the attack. Each decision maker assesses differently the probability of the result of an attack, which depends on the defense and attack adopted: $p_D(s \mid d,a)$ and $p_A(s \mid d,a)$. The utility function of the Defender $u_D(d,s)$ depends on her chosen defense and the result of the attack. Similarly, the Attacker's utility function is $u_A(a,s)$. We first recall the standard game theoretic approach and check its robustness. We then present the ARA solution and, again, provide a robust analysis.

3.3.1 Game Theoretic Solution and Robustness

The standard game theoretic solution does not require the Attacker to know the Defender's probabilities and utilities, since he observes the Defender's actions. However, the Defender needs to know the Attacker's utilities and probabilities (u_A, p_A), an example of common knowledge. We then proceed as follows. First, we compute the expected utilities of the players at node S in Fig. 3.1:

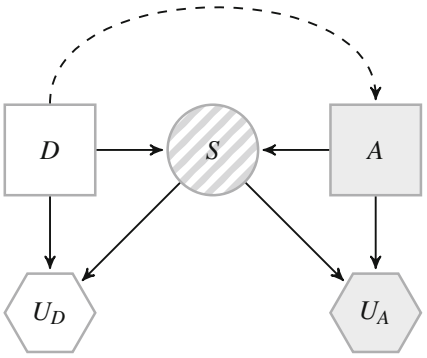

Fig. 3.1: The two player sequential decision game

$$\psi_A(a,d) = \int u_A(a,s)p_A(s|a,d)ds, \tag{3.1}$$

$$\psi_D(a,d) = \int u_D(d,s)p_D(s|a,d)ds.$$

Then, we compute the Attacker's best response to the Defender's action d, which is

$$a^*(d) = \text{argmax}_{a\in\mathscr{A}}\psi_A(a,d).$$

Knowing this, the Defender's optimal action is, then,

$$d^*_{GT} = \text{argmax}_{d\in\mathscr{D}}\psi_D(a^*(d),d).$$

The solution $\left(a^*(d^*_{GT}), d^*_{GT}\right)$ is a Nash equilibrium and, indeed, a sub-game perfect equilibrium, see [5]. We call d^*_{GT} the Nash defense.

3.3.1.1 Robustness of the Game Theoretic Solution

Since we are supporting the Defender, we could argue that we know reasonably well (u_D, p_D). However, we would contend that knowledge about (u_A, p_A) is that precise, since it would require the Attacker to reveal them (common knowledge). This is questionable in many application areas including security, cybersecurity and competitive marketing. We may use robust methods to criticize such information and, consequently, assess the game theoretic solution.

As discussed in Sect. 3.2, from a conceptual point of view, to perform robustness we may consider classes for the Attacker's utilities and probabilities that we model through $u \in \mathscr{U}_A$, $p \in \mathscr{P}_A$. Then, mimicking the approach above, for each feasible (u,p) we could:

- Compute the expected utilities $(\psi_A^{u,p}(d,a), \psi_D^{u,p}(d,a))$ at node S in Fig. 3.1.
- Compute the best response attack $a^*_{u,p}(d)$ for each d.
- Compute the optimal defense $d^*_{u,p}$.

Then, if $d^*_{u,p}$ remains reasonably stable for the allowed perturbations of u and p, with $u \in \mathscr{U}_A$, $p \in \mathscr{P}_A$, the game theoretic solution seems robust. However, if $d^*_{u,p}$ is not that stable, we have an issue which questions, at first sight, the relevance of the proposed Nash defense d^*_{GT}. At a deeper level, it also questions the appropriateness of the (u_A, p_A) assessment, actually serving to criticize the game theoretic assumptions, specially that of common knowledge, see [16] or [10].

From an operational point of view, the above robustness analysis scheme for the game theoretic approach boils down to two computational issues:

- Exploring the whole range of perturbations $u \in \mathscr{U}_A$, $p \in \mathscr{P}_A$. In some cases, for classes of probabilities and utilities widely studied in the robust Bayesian literature, see [2], it is possible to identify the extremal elements of $\mathscr{U}_A$ and $\mathscr{P}_A$ and compute upper and lower bounds on the quantities of interest (namely optimal decisions $d^*_{u,p}$ and their expected utilities), through numerical optimization methods. Another possible approach would be to randomly sample elements u, p from the sets $\mathscr{U}_A, \mathscr{P}_A$ and check for eventual large variations in $d^*_{u,p}$ (and their expected utilities).
- Declaring whether the effects induced by changes over $d^*_{u,p}$ and the expected utility are sufficiently small. As discussed in Sect. 3.2, a possible criterion could be given by the range spanned by $d^*_{u,p}$ as utility and probability vary in the classes, i.e. $u \in \mathscr{U}_A$ and $p \in \mathscr{P}_A$, respectively. Regarding the effects on the expected utility, a criterion of interest could be based on the regret $r_{u,p}(d^*_{GT})$ given by the difference in expected utility when considering, for a given pair (u, p), the Nash defense d^*_{GT} and the optimal defense $d^*_{u,p}$. A small value of $\sup_{(u,p)\in\mathscr{U}_A\times\mathscr{P}_A} r_{u,p}(d^*_{GT})$ would denote robustness with respect to the choice of utility and probability and, therefore, any pair (u, p) can be chosen as opinion on the Attacker's behavior with no significant change in the consequences. If robustness is not achieved, then we could undertake a minimum regret approach as discussed in Sect. 3.2.

An alternative would be to move to ARA, as discussed next.

3.3.2 ARA Solution and Robustness

If the game theoretic solution is not robust, then we need to address the issue. One way forward is to perform an ARA approach. For this, we weaken the common knowledge assumption. In the sequential game, this means that the Defender does not know (p_A, u_A). The problem she faces is depicted in Fig. 3.2.

To solve her problem, the Defender requires more information than $p_D(s|a,d)$ and $u_D(d,s)$, available from our earlier discussion. She also needs $p_D(a|d)$, which is her assessment of the probability that the Attacker will choose attack a after having observed that she has chosen the defense d. Once the Defender has completed these assessments, she can solve the problem. Indeed, the expected utility of d would be

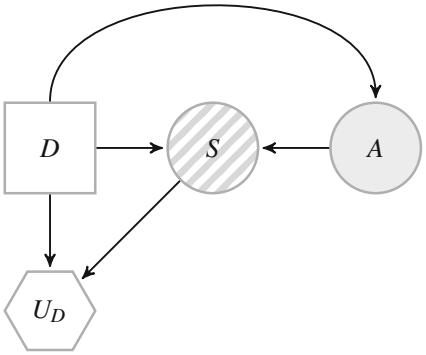

Fig. 3.2: The decision problem as seen by Defender

$$\psi_D(d) = \int \psi_D(a,d) p_D(a|d) da = \int \left[\int u_D(d,s) p_D(s|a,d) ds \right] p_D(a|d) da.$$

Finally, her optimal decision would be $d^*_{ARA} = \text{argmax}_{d \in \mathscr{D}} \psi_D(d)$. Note that, in terms of classic game theory, the solution d^*_{ARA} for our sequential game may not correspond to a Nash equilibrium, see the example in Sect. 3.5.

Eliciting $p_D(a|d)$ requires the Defender to analyze the problem from the Attacker's perspective.

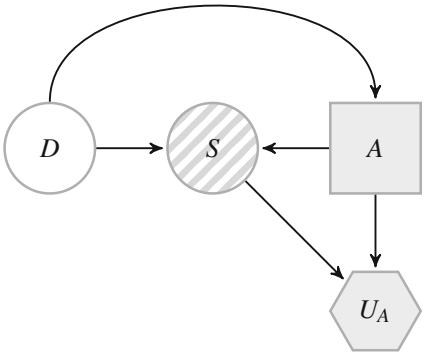

Fig. 3.3: Defender's analysis of Attacker's problem

First, the Defender puts herself in the Attacker's shoes, and thinks about his decision problem. Figure 3.3 represents the Attacker's problem, as seen by the Defender, assuming he is an expected utility maximizer. The Defender will use all the information and judgment that she can obtain about the Attacker's utilities and probabilities. Instead of using point estimates for p_A and u_A to find the Attacker's optimal decision $a^*(d)$ for a given d, the Defender's uncertainty about the Attacker's decision should derive from her uncertainty about the Attacker's (p_A, u_A), through a distribution F

on the space of utilities and probabilities, which we designate random probabilities and utilities. This induces a distribution over the Attacker's expected utility in (3.1), where the random expected utility for A would be

$$\Psi_A(a,d) = \int U_A(a,s)P_A(s|a,d)ds,$$

for $(P_A, U_A) \sim F$. Then, the Defender would find

$$p_D(a|d) = \mathbb{P}_F[a = \text{argmax}_{x\in\mathscr{A}}\Psi_A(x,d)],$$

in the discrete case and, similarly, in the continuous case. We can use Monte Carlo simulation to approximate $p_D(a|d)$ by drawing N samples $\left\{\left(P_A^i, U_A^i\right)\right\}_{i=1}^N$ from F and setting

$$\hat{p}_D(a|d) \approx \frac{\#\{a = \text{argmax}_{x\in\mathscr{A}}\ \Psi_A^i(x,d)\}}{N}, \tag{3.2}$$

where $\Psi_A^i(a,d) = \int U_A^i(a,s)P_A^i(s|a,d)ds$.

3.3.2.1 Robust Analysis

The above approach leads to a Bayesian decision analysis problem with the peculiarity that we have a complex procedure to forecast the adversarial actions. To do so, we formulate the adversary decision making problem and propagate our uncertainty about the adversary judgments to the optimal adversarial action.

We could then think about performing a robust Bayesian analysis. The inputs to the Defender's decision analysis are $(u_D(d,s), p_D(s|a,d), p_D(a|d))$. We focus here on sensitivity to the last component $p_D(a|d)$, surely the most contentious one, attained through adversarial calculations based on the proposed $U_A(a,s), P_A(s|a,d)$. For that, we define classes $\mathscr{U}_A$, $\mathscr{P}_A$ of random utilities and probabilities. For each pair U, P in such class, we define $p_D^{UP}(a|d)$ through the ARA approach which, in turn, leads to d_{ARA}^{*UP}.

Then, it is possible to consider the impact of the imprecision about U and P over three quantities: $p_D^{UP}(a|d)$, d_{ARA}^{*UP} and $\psi(d_{ARA}^{*UP})$. The first quantity requires the comparison of densities (actually of their Monte Carlo approximations) using indices like the Kullback-Leibler divergence or Gini index. For the first and second quantities, the interest centers around the variation of the decision (for the Defender), whereas for the third one, the focus is on the expected utility of the decision. The last quantity should be of major interest. In all three cases, we say that robustness holds when the value of interest does not change much, whereas additional analysis should be taken otherwise, as described in Sect. 3.2. In particular, if the distributions $p_D^{UP}(a|d)$ do not differ too much, it is possible to choose one of them and use d_{ARA}^{*UP} directly.

3.3.3 A Full Robust Solution

If the ARA analysis is not robust, we may opt for gathering additional information to reduce the classes $\mathscr{U}_A$ and $\mathscr{P}_A$. The choice of increasing the sample size in the Monte Carlo estimation $\hat{p}_D(a|d)$ in (3.2) would be useful in reducing the variability of the distribution. However, it will typically be ineffective in increasing robustness.

Once all possible sources of information have been exploited to try to increase robustness about d^*_{ARA} and $\psi(d^*_{ARA})$, then some extra criterion has to be introduced to make a decision and report a value about the quantity of interest. In any case, such decision should be reported with the warning of lack of robustness. As discussed in Sect. 3.2, we could consider the decision d^*_R minimizing the maximum regret, i.e.

$$\min_d \max_{U\in U_A, P\in P_A} \left[\int \psi_D(a, d^{*UP}_{ARA}) p^{UP}_D(a|d^{*UP}_{ARA}) da - \int \psi_D(a,d) p^{UP}_D(a|d) da \right].$$

3.4 Simultaneous Games

We discuss now the simultaneous game model: two agents choose their decisions, without knowing the action selected by each other. Among others, see [27] for a related discussion within a game theoretic framework. As an example, imagine a case in which the EASA decides whether to introduce undercover marshals in an airplane that might, or not, be hijacked by terrorists.

Assume that the adversaries have alternative sets $\mathscr{D}$ and $\mathscr{A}$ of defenses and attacks, respectively. The only relevant uncertainty is S, denoting the success s of the attack. Each decision maker assesses differently the probability of the result of the attack, which depends on the defense and attack adopted: $p_D(s \mid d,a)$ and $p_A(s \mid d,a)$. The utility function of the Defender $u_D(d,s)$ depends on her chosen defense and the result of the attack. Similarly, the Attacker's utility function is $u_A(a,s)$, as illustrated in Fig. 3.4.

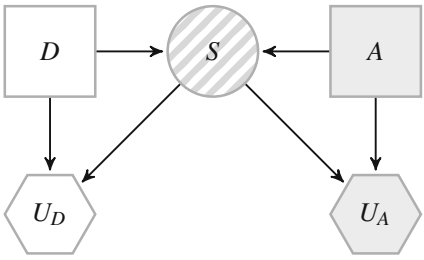

Fig. 3.4: BAID for the simultaneous Defend-Attack model

3.4.1 Game Theoretic Solution

Under common knowledge, preferences and beliefs from both the Defender and the Attacker, (u_D, p_D) and (u_A, p_A) respectively, are disclosed. Therefore, each adversary knows the expected utility that each pair $(d,a) \in \mathscr{D} \times \mathscr{A}$ would provide to both of them, computed through

$$\psi_D(d,a) = \int u_D(d,s) p_D(s|a,d) ds,$$

$$\psi_A(d,a) = \int u_A(a,s) p_A(s|a,d) ds.$$

A Nash equilibrium (d^*_{GT}, a^*_{GT}) for this game would satisfy

$$\psi_D(d^*_{GT}, a^*_{GT}) \geq \psi_D(d, a^*_{GT}) \ \forall d \in \mathscr{D} \quad \text{and}$$
$$\psi_A(d^*_{GT}, a^*_{GT}) \geq \psi_A(d^*_{GT}, a) \ \forall a \in \mathscr{A}.$$

Finding Nash equilibria may require the use of randomized strategies, see [4]. There could be several equilibria, with no unambiguous criteria to further discern among them, see [16] for a discussion.

If utilities and probabilities are not common knowledge among the adversaries, a game-theoretic approach proceeds by modeling the game as one with incomplete information, see [6–8], by introducing the notion of player types. Each player will be of a certain type which is known to him but not to his opponent: a player's type represents the private information he may have. Harsanyi proposes the Bayes-Nash equilibrium as a solution concept, still under a strong common knowledge assumption: the adversaries' beliefs about the opponent's types are common knowledge and modeled through a common prior distribution. Moreover, it is assumed that the players' beliefs about other uncertainties in the problem are also common knowledge. Again randomized strategies might be required to find such equilibria.

3.4.1.1 Robustness of the Game Theoretic Solution

We could argue that we know reasonably well (u_D, p_D), since we are supporting the Defender. However, we would contend that (u_A, p_A) is properly known, since it requires common knowledge, which is questionable. To address this concern, we perform a robust analysis of the Defender's decision at the Nash equilibrium.

For that, we would consider classes for the Attacker's utilities and probabilities represented as $u \in \mathscr{U}_A$, $p \in \mathscr{P}_A$. Then, for each feasible (u,p) we could compute the corresponding Nash equilibrium (d^*_{up}, a^*_{up}). If d^*_{up} remains stable for the feasible perturbations of u and p, the game theoretic solution d^*_{GT} seems robust, from the perspective of the Defender. However, if d^*_{up} changes, specially the corresponding expected utility, we have a problem which questions, at first sight, the relevance of the proposed d^*_{GT} and, at a deeper level, the appropriateness of the (u_A, p_A) assessment,

actually serving to criticize the game theoretic approach at large and, in particular, the common knowledge assumption. The two computational issues about finding all possible optimal decisions and assessing robustness are dealt with as mentioned in Sect. 3.3.3.

Note that we could actually study robustness with respect to (u_D, p_D, u_A, p_A) and consider changes in $d^*_{u_A,p_A,u_D,p_D}$. In this case, if the Defender's Nash equilibrium decision is sensitive, we might question the Defender's knowledge, besides the game theory postulates.

3.4.2 ARA Solution and Robustness

If the Nash equilibrium is unstable, we may try an ARA approach. We have to weaken the common (prior) knowledge assumptions. As reflected in Fig. 3.5, the Defender has to choose a defense $d \in \mathscr{D}$, whose consequences depend on the success of an attack $a \in \mathscr{A}$ simultaneously chosen by the Attacker, which is, therefore, uncertain for the Defender at the time she makes her decision.

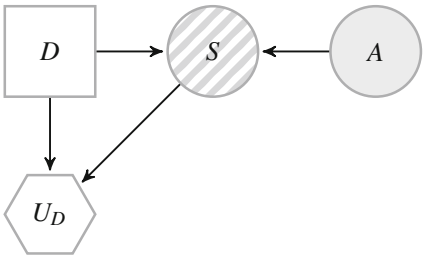

Fig. 3.5: The Defender's decision analysis

By standard Decision Theory, the Defender should maximize her expected utility, see [3]. The Defender knows her utility function $u_D(d,s)$ and her probability assessment p_D over S, conditional on (d,a). However, she does not know the Attacker's decision a at node A. She expresses her uncertainty through a probability distribution $p_D(a)$. Then, the optimization problem she should solve to find d^*_{ARA} is

$$\begin{aligned} \max_d \int \psi_D(a,d) p_D(a) da &= \max_d \int [\int u_D(d,s) p_D(s|a,d)\, ds]\, p_D(a)\, da \\ &= \max_d \int \int u_D(d,s) p_D(s|a,d) p_D(a)\, ds\, da. \end{aligned} \tag{3.3}$$

We could then perform a robust analysis based on $u_D, p_D(s|a,d)$ and $p_D(a)$. However, eliciting this last probability distribution is more difficult. We may use ARA as follows to get it.

Suppose the Defender thinks that the Attacker is an expected utility maximizer who tries to solve the decision problem shown in Fig. 3.6. The Attacker would look for the attack $a \in \mathscr{A}$ providing him maximum expected utility:

$$a^* = \arg\max_{a \in \mathscr{A}} \int \int u_A(a,s) p_A(s|a) p_A(d) ds\, dd.$$

In general, the Defender will be uncertain about the Attacker's utility function and probabilities, and she would consider random utilities and probabilities through $F = (U_A(a,s), P_A(s|a), P_A(d))$ and compute the random optimal alternative

$$A^*|D = \arg\max_{a \in \mathscr{A}} \int \int U_A(a,s) P_A(s|a) P_A(d) ds\, dd. \tag{3.4}$$

Then, we would make

$$p_D(a) = P(A^* = a|D)$$

in the discrete case and, similarly, in the continuous case.

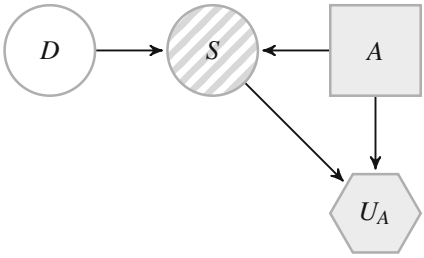

Fig. 3.6: The Attacker's decision analysis, as seen by the Defender

Note that $(U_A(a,s), P_A(s|a))$ would be comparatively easily elicited from the Defender, see examples in [1]. However, the elicitation of $P_A(d)$ may require further analysis leading to a next level of recursive thinking: the Defender would need to think about how the Attacker analyzes her problem. This is why we condition in (3.4) by (the distribution of) D.

In the above, the Defender presumes that the Attacker thinks she is an expected utility maximizer trying to solve a decision problem like that described in Fig. 3.5. Therefore, in order for the Defender to assess the required distribution, she will elicit (U_A, P_A) from her viewpoint, and assess $P_A(D)$ through the analysis of her decision problem, as thought by the Attacker, mimicking the resolution of problem (3.3) from the Attacker's perspective. This reduces the assessment of $P_A(D)$ to computing the distribution

$$D \mid A^1 \sim \arg\max_{d \in \mathscr{D}} \int \int U_D(d,s) P_D(S = s \mid d,a) P_D(A^1 = a) ds\, da,$$

assuming that the Defender is able to assess $P_D(A^1)$. A^1 represents the Attacker's decision within the Defender's second level of recursive thinking in the nested decision model used by the Defender to predict the Attacker's analysis of her decision problem. To complete the assessment, the Defender should elicit $(U_D, P_D) \sim G$, representing her probabilistic knowledge about how the Attacker may estimate her

utility function $u_D(d,s)$ and her probability p_D over $S|d,a$, when she analyzes how the Attacker thinks about her decision problem. The elicitation of $P_D(A^1)$ might require further recursive thinking from the Defender, see our final discussion.

3.4.2.1 Robustness

Performing a robust analysis for the ARA approach to the simultaneous game would be similar to what described earlier. Consider a class for $(U_A(a,s),P_A(s|a),P_A(d)) \in (\mathscr{U}_A,\mathscr{P}_A,\mathscr{Q}_A)$. We use (U,P,Q) to simplify the notation describing the elements in the classes. Then, for (U,P,Q) satisfying the constraints, replicating the approach above we could compute p_D^{UPQ} and d_{ARA}^{*UPQ}. If d_{ARA}^{*UPQ} remains stable with respect to changes in (U,P,Q), then the problem seems robust and we could apply the ARA approach with little concern. Otherwise, we could still use a robust solution concept, like the minimum regret mentioned in Sect. 3.2.

3.5 An Example

As an illustration, we consider a sequential defend-attack security routing problem. An organization needs to make a trip, either through a safe, but costly, route, or through a cheaper, but more dangerous, route. In this case, they may invest in security, rendering the route less dangerous. See [23] for a case concerning piracy in Somalia. Table 3.1 displays the consequences, expressed as costs, for various defend and attack possibilities.

Table 3.1: Loss function in routing problem

Defense	Attack	Attack result	Def. cons.	Att. cons.
Dang. prot	Attack	θ_1	$c\theta_1+K$	$-d\theta_1+B$
	No Attack		K	0
Dang. unp	Attack	θ_2	$c\theta_2$	$-d\theta_2+B$
	No attack		0	0
Safe			H	0

The following parameters are used:

- θ_1 represents the fraction of assets lost by the organization when attacked but protected.
- θ_2 represents the fraction of assets lost by the organization when attacked and not protected.
- c is the cost per unit of assets.
- K are the protection costs.
- H is the cost of going through the expensive route.

- d is the Attacker's gain per unit of assets lost by the Defender.
- B is the cost of an attack.

The Defender has beliefs for θ_i, with $\theta_i \sim \beta(a_i, b_i), i = 1, 2$. She is risk averse and her utility function is strategically equivalent to $-\exp(hx)$, where x is her cost and $h > 0$ is her risk aversion coefficient. The Attacker has different beliefs for θ_i with $\theta_i \sim \beta(c_i, e_i), i = 1, 2$. He is risk prone and his utility function is strategically equivalent to $\exp(-mx)$, where x is his cost and $m > 0$ is his risk proneness coefficient. Both agents expect θ_1 to be smaller than θ_2, but not necessarily. This may be reflected in the choice of the beta parameters, for example with $a_1/(a_1 + b_1) < a_2/(a_2 + b_2)$, in the case of the Defender. Table 3.2 provides the expected utilities for both agents under various interaction scenarios.

Table 3.2: Expected utilities in routing problem

Interaction	Eu. def	Eu. att
Prot, Att.	$-\int e^{h(c\theta_1+K)} f(\theta_1 \mid a_1, b_1) d\theta_1$	$\int e^{m(d\theta_1-B)} f(\theta_1 \mid c_1, e_1) d\theta_1$
Prot.,NoAtt.	$-e^{hK}$	1
NoProt.,Att.	$-\int e^{h(c\theta_2)} f(\theta_2 \mid a_2, b_2) d\theta_2$	$\int e^{m(d\theta_2-B)} f(\theta_2 \mid c_2, e_2) d\theta_2$
NoProt.,NoAtt.	-1	1
Safe	$-e^{hH}$	1

The problem may be viewed through the game tree in Fig. 3.7, where d_1 means going through the dangerous route but protected; d_2 means going through the dangerous route but unprotected; and, finally, d_3 means going through the safe route, whereas a means *attack* and $\bar{a}$ means *no attack*.

We are supporting the Defender and assess from her the values $c = 200{,}000$, $K = 50{,}000$, $H = 100{,}000$, $h = 3$. We also elicit from her the distributions $\beta(a_1, b_1)$, with mean 0.3 and standard deviation 0.07, leading to $a_1 = 12.325$, $b_1 = 28.76$; and $\beta(a_2, b_2)$, with mean 0.7 and standard deviation 0.18, leading to $a_2 = 3.815$, $b_2 = 1.635$.

3.5.1 Game Theoretic Approach

Under common knowledge, we assume the Defender knows that $d = 30{,}000$, $B = 10{,}000$, $m = 5$ and the distributions $\beta(c_1, e_1)$, with mean 0.313 and standard deviation 0.16, leading to $c_1 = 2.272$, $e_1 = 4.978$; and $\beta(c_2, e_2)$, with mean 0.324 and standard deviation 0.11, leading to $c_2 = 5.49$, $e_2 = 11.45$. We, then, proceed as follows:

- At node A_1, compute $\max(\psi_A(d_1, a), \psi_A(d_1, \bar{a}))$ and call the optimal action $a^*(d_1)$. In the example, we have $\max(1.001, 1) = 1.001$ and the optimal decision for the Attacker is a.

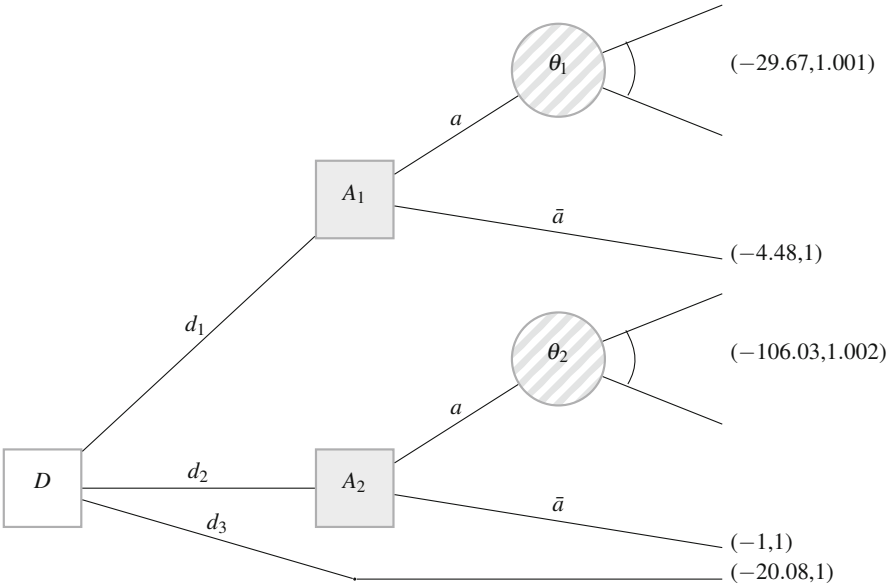

Fig. 3.7: Game tree for the routing problem

- At node A_2, compute max $(\psi_A(d_2,a),\psi_A(d_2,\bar{a}))$ and call the corresponding action $a^*(d_2)$. We have max $(1.002,1) = 1.002$ and the optimal decision for the Attacker is a.
- At node D, compute max $(\psi_D(d_1,a^*(d_1)),\psi_D(d_2,a^*(d_2)),\psi_D(d_3))$ and call the optimal action d^*_{GT}. In our case, max $(-29.67,-106.03,-20.08) = -20.08$ and the Nash defense d^*_{GT} is d_3, that is, to choose the safe route.

3.5.2 Robustness of the Game Theoretic Solution

We consider now the robustness of the game theoretic solution. We simplify and assume that the attack cost $B = 10{,}000$ is reasonably well known. Assume that d is not that well known and we express this through a constraint $d \in [10000, 50000]$. Similarly, suppose that $c_1 \in [0,3]$, $e_1 \in [1,6]$, $c_2 \in [2,8]$ and $e_2 \in [10,14]$. We sample randomly from these intervals 1000 times and repeat the procedure in Sect. 3.5.1.

The three defenses may be Nash, given the constraints. Indeed, based on the above sampling scheme, we estimate that the probabilities of the three alternatives being Nash are, respectively, 0.454, 0.236 and 0.31, therefore with no clear winner. The maximum loss when we implement the defense $d^*_{GT} = d_3$ is 19.08. This is deemed large enough and we need to perform an ARA approach.

3.5.3 ARA Approach

The problem faced by the Defender is described in the decision tree in Fig. 3.8.

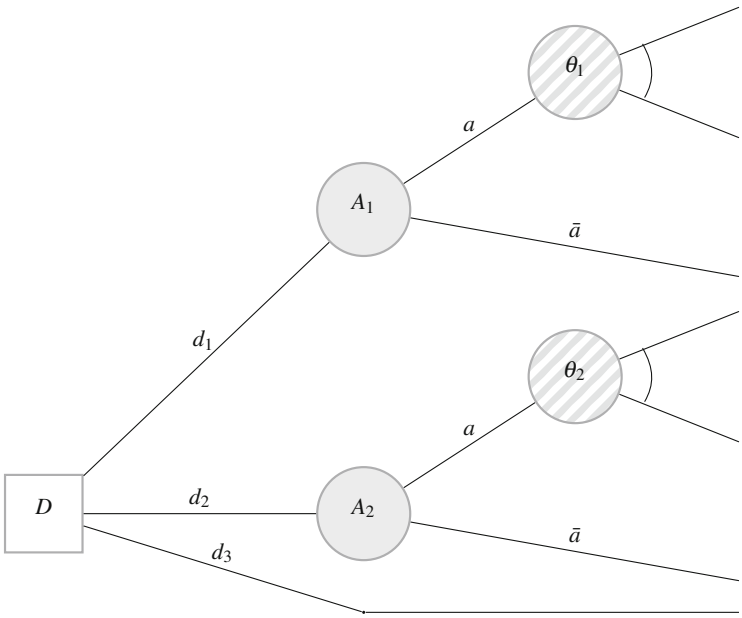

Fig. 3.8: Decision tree for the Defender in the routing problem

The expected utilities of the first two alternatives have the form

$$\psi_D(d_i) = p_D(a|d_i)\psi_D(d_i,a) + p_D(\bar{a}|d_i)\psi_D(d_i,\bar{a}), i = 1,2.$$

Thus, we need to assess the attack probabilities $p(a|d_i)$ given the implemented defense d_i.

We illustrate the estimation of $p_D(a|d_1)$. We assume that d, c_1, e_1, c_2, e_2 are uniformly distributed over the intervals described in Sect. 3.5.2. Thus, we assume that $d \sim \mathscr{U}[10000, 50000]$, $c_1 \sim \mathscr{U}[0,3]$, $e_1 \sim \mathscr{U}[1,6]$, $c_2 \sim \mathscr{U}[2,8]$ and $e_2 \sim \mathscr{U}[10,14]$. Then, we may use Algorithm 1 to estimate the required probability, where $\psi_A^k(d_1,x)$ designates the expected utility that the Attacker reaches, when the Defender implements d_1 and he implements attack x and the sampled parameters are $d^k, c_1^k, c_2^k, e_1^k, e_2^k$.

In our particular case, with $N = 10{,}000$, we obtain $\hat{p}(a|d_1) = 0.406$ (and, consequently, $\hat{p}(\bar{a}|d_1) = 0.594$). Similarly, $\hat{p}(a|d_2) = 0.764$ and $\hat{p}(\bar{a}|d_2) = 0.236$. Then, we have $\psi(d_1) = -14.7$, $\psi(d_2) = -81.2$ and $\psi(d_3) = -20.08$ and the optimal ARA defense d^*_{ARA} is d_1, which is different to d^*_{GT}.

Algorithm 1: Estimating $p(a|d_1)$

```
p = 0;
for k ← 1 to N do
    Sample d^k, c_1^k, c_2^k, e_1^k, e_2^k;
    if ψ_A^k(d_1, a) ≥ ψ_A^k(d_1, ā) then
        p = p + 1;
p̂(a|d_1) = p/N;
```

3.5.4 Robustness of the ARA Solution

We consider now the robustness of the ARA solution. For that, we consider classes of beta distributions with the same support than the corresponding parameters. As an example, for d, we shall assume that $d \sim \beta[o_1, o_2]$ over the interval $[10000, 50000]$, with $o_1 \in [0.5, 1.5]$, $o_2 \in [0.5, 1.5]$. Similarly, for the other parameters we use beta distributions over the previous intervals, with parameters as in Table 3.3, where the first parameter of the beta distribution is uniform over [LL, LU] and the second parameter of the beta distribution is uniform over [UL, UU].

We sample 100 times from such distributions and repeat the procedure in Sect. 3.5.3. Then, the estimated probabilities of each defense being optimal, in the ARA sense, would be, respectively, $\hat{p}(d_1) = 1$, $\hat{p}(d_2) = 0$ and $\hat{p}(d_3) = 0$. Therefore, d_1 seems clearly the most likely alternative for being optimal.

The regrets when we implement various solutions, are respectively, 0 for d_1, 37.91 for d_2 and 8.54 for d_3. Thus, the minimum regret defense is d_1.

Table 3.3: Upper and lower limits for the parameters of the involved beta distributions

Parameter	LL	LU	UL	UU
c_1	0.5	1.5	0.5	1.5
c_2	0.5	1.5	0.5	1.5
e_1	0.5	1.5	0.5	1.5
e_2	0.5	1.5	0.5	1.5

3.6 Discussion

Adversarial Risk Analysis is an emergent paradigm when supporting a decision maker who faces adversaries and such that the consequences are random and depend on the actions of all participating agents. The prevalent paradigm in this area is Game Theory. In this chapter, we have provided a framework for robustness analysis in this area.

The approach we have followed is:

- Under common knowledge assumptions compute the game theoretic solution. Perform a robust analysis for such solution. If it is stable, such solution may be used with confidence and we do not require further analysis.
- Otherwise, perform an ARA. Undertake a robust analysis for the ARA solution. If it is stable, the ARA solution may be used with confidence and the analysis stops. Otherwise, gather more data and/or refine the relevant classes, eventually declaring the robustness of the ARA solution. If not sufficient, move towards next stage.
- Undertake a minimum regret (or other robust) concept.

We have illustrated it with two simple models, the sequential defend-attack and the simultaneous defend-attack, but the ideas would extend to more complex ARA models. Similarly, we have assumed that the attacker was maximising expected utility but the ideas may be translated to other attacker rationalities, as in [21].

There are many other sensitivity analysis questions relevant in ARA. For example, we mentioned above the recursive assessment required in the simultaneous game, which may be expressed as follows, see [17]:

Algorithm 2: Recursive assessment required in the simultaneous game

for $i \leftarrow 1$ **to** ∞ **do**

Find $\Pi_{D^{i-1}}(A^i)$ by solving

$$A^i \mid D^i \sim \arg\max_{a \in \mathscr{A}} \sum_{d \in \mathscr{D}} \left[\sum_{s \in \{0,1\}} U_A^i(a,s) \;\; P_A^i(S = s \mid d, a) \right] \Pi_{A^i}(D^i = d)$$

with $(U_A^i, P_A^i) \sim F^i$

Find $\Pi_{A^i}(D^i)$ by solving

$$D^i \mid A^{i+1} \sim \arg\max_{d \in \mathscr{D}} \sum_{a \in \mathscr{A}} \left[\sum_{s \in \{0,1\}} U_D^i(d,s) \;\; P_D^i(S = s \mid d, a) \right] \Pi_{D^i}(A^{i+1} = a)$$

with $(U_D^i, P_D^i) \sim G^i$

$i = i + 1$;

This hierarchy would stop when the Defender lacks the information necessary to assess the distribution F^i or G^i associated with the decision analysis of A^i and D^i, respectively. At this point, the Defender would assign an unconditional probability distribution over A^i or D^i, respectively, without going deeper into the hierarchy, summarizing all the information she might have through the direct assessment of $\Pi_{D^{i-1}}(A^i)$ or $\Pi_{A^i}(D^i)$, as might correspond. Should she have no additional information to do so, she could assign a noninformative distribution, see [3].

However, climbing up one level in the hierarchy entails a lot of effort. We could question whether this is worth it by using value of information types of computation.

Acknowledgements

DRI, CA and JG are grateful to the MINECO MTM2014-56949-C3-1-R and the AXA-ICMAT Chair in Adversarial Risk Analysis. The authors are grateful to the support of the COST IS1304 Action on Expert Judgement.

References

1. Banks, D.L., Aliaga, J.M.R., Ríos Insua, D.: Adversarial Risk Analysis, vol. 343. CRC Press, Boca Raton, FL (2015)
2. Berger, J.O., Ríos Insua, D., Ruggeri, F.: Bayesian robustness. In: Ríos Insua, D., Ruggeri, F. (eds.) Robust Bayesian Analysis, pp. 1–32. Springer, New York (2000)
3. French, S., Ríos Insua, D.: Statistical Decision Theory. Wiley, New York (2000)
4. Gibbons, R.: A Primer in Game Theory. Harvester Wheatsheaf, Hemel Hempstead (1992)
5. Hargreaves-Heap, S., Varoufakis, Y.: Game Theory: A Critical Introduction. Routledge, Abingdon (2004)
6. Harsanyi, J.C.: Games with incomplete information played by "bayesian" players, part i. The basic model. Manag. Sci. **14**(3), 159–182 (1967)
7. Harsanyi, J.C.: Games with incomplete information played by "bayesian" players, part ii. Bayesian equilibrium points. Manag. Sci. **14**(5), 320–334 (1968)
8. Harsanyi, J.C.: Games with incomplete information played by "bayesian" players, part iii. The basic probability distribution of the game. Manag. Sci. **14**(7), 486–502 (1968)
9. Kardes, E.: Robust Stochastic Games and Applications to Counter-Terrorism Strategies. Center for Risk and Economic Analysis of Terrorism Events, University of Southern California, Los Angeles (2005)
10. Lippman, S.A., McCardle, K.F.: Embedded Nash bargaining: risk aversion and impatience. Decis. Anal. **9**(1), 31–40 (2012)
11. McLay, L., Rothschild, C., Guikema, S.: Robust adversarial risk analysis: a level-k approach. Decis. Anal. **9**(1), 41–54 (2012)
12. Merrick, J., Parnell, G.S.: A comparative analysis of PRA and intelligent adversary methods for counterterrorism risk management. Risk Anal. **31**(9), 1488–1510 (2011)
13. Nau, R.F.: Joint coherence in games of incomplete information. Manag. Sci. **38**(3), 374–387 (1992)
14. O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T.: Uncertain Judgements: Eliciting Experts' Probabilities. Wiley, Hoboken, NJ (2006)
15. Parnell, G., Banks, D., Borio, L., Brown, G., Cox, L., Gannon, J., Harvill, E., Kunreuther, H., Morse, S., Pappaioanou, M., et al.: Report on methodological improvements to the department of homeland security's biological agent risk analysis National Academies Press, Washington DC (2008)
16. Raiffa, H., Richardson, J., Metcalfe, D.: Negotiation Analysis: The Science and Art of Collaborative Decision Making. Harvard University Press, Cambridge, MA (2002)
17. Rios, J., Ríos Insua, D.: Adversarial risk analysis for counterterrorism modeling. Risk Anal. **32**(5), 894–915 (2012)
18. Ríos Insua, D., Ruggeri, F.: Robust Bayesian Analysis. Springer, New York (2000)

19. Ríos Insua, D., Ruggeri, F., Vidakovic, B.: Some results on posterior regret γ-minimax estimation. Stat. Decis. **13**, 315–331 (1995)
20. Ríos Insua, D., Rios, J., Banks, D.: Adversarial risk analysis. J. Am. Stat. Assoc. **104**(486), 841–854 (2009)
21. Ríos Insua, D., Banks, D., Rios, J.: Modeling opponents in adversarial risk analysis. Risk Anal. (2015)
22. Rothschild, C., McLay, L., Guikema, S.: Adversarial risk analysis with incomplete information: a level-k approach. Risk Anal. **32**(7), 1219–1231 (2012)
23. Sevillano, J.C., Rios Insua, D., Rios, J.: Adversarial risk analysis: the Somali pirates case. Decis. Anal. **9**(2), 86–95 (2012)
24. Stahl, D.O., Wilson, P.W.: On players models of other players: theory and experimental evidence. Games Econ. Behav. **10**(1), 218–254 (1995)
25. Von Winterfeldt, D., O'Sullivan, T.M.: Should we protect commercial airplanes against surface-to-air missile attacks by terrorists? Decis. Anal. **3**(2), 63–75 (2006)
26. Wang, C., Bier, V.M.: Expert elicitation of adversary preferences using ordinal judgments. Oper. Res. **61**(2), 372–385 (2013)
27. Zhuang, J., Bier, V.M.: Balancing terrorism and natural disasters-defensive strategy with endogenous attacker effort. Oper. Res. **55**(5), 976–991 (2007)